

Adaptive Operator Selection with Dynamic Multi-Armed Bandits

L. Da Costa¹, A. Fialho², M. Schoenauer^{1,2}, M. Sebag^{1,2}

¹Team TAO, LRI – tao.lri.fr
INRIA Saclay
FRANCE

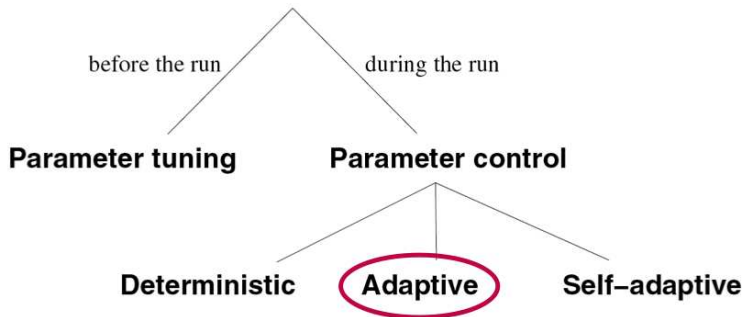
²Microsoft Research-INRIA Joint Centre
FRANCE

GECCO 2008 – July 16th.

Parameter Control in EAs

Best optimization methods require specific developments in order to "cross the chasm" – "Self-tuning" methods?

Parameter Setting in EAs



(from Eiben *et al.*, in *Parameter Setting in EAs* (2007), pages. 19–46.)

Adaptive Operator Selection: an example

One-Max Problem

- Representation: bit-string
- Fitness: number of 1's in individual

$(1, \lambda)$ -EA for the One-Max Problem

No crossover, 2 mutation operators:

- 5-bit flip
- bit-flip

Q: Which operator is better?

A: It depends on "where we are"

- | | | |
|---|---|---|
| 0 | 0 | 0 |
|---|---|---|

 ...

0	0	0	0	0
---	---	---	---	---
- | | | |
|---|---|---|
| 1 | 1 | 1 |
|---|---|---|

 ...

1	1	1	1	0
---	---	---	---	---

Stochastic K-armed "bandit"

- Each arm j described by a probability p_j
- On $t = 1, \dots, T$, gambler plays arm j

$$\text{reward at } t : r_t = \begin{cases} 1 & , \text{with prob} = p_j \\ 0 & , \text{with prob} = 1 - p_j \end{cases}$$

- Goal: maximize *total cumulated reward* (**TCR**) or minimize *regret* (*loss*):

$$\max \sum_{t=1}^T r_t \equiv \min \mathcal{L}(T), \text{ where } \mathcal{L}(T) = T * p^* - \sum_{t=1}^T r_t$$

Q: What is the best strategy for a gambler?

Multi-Armed Bandits: a little bit of theory

- Optimal regret: $\mathcal{L}(T) = \mathcal{O}(\log T)$
- **UCB1**¹: at t , choose arm j maximizing :

$$\hat{r}_{j,t} + \sqrt{\frac{2 \log \sum_k n_{k,t}}{n_{j,t}}}, \text{ where } \begin{cases} \hat{r}_{j,t} & \text{, estimated reward for arm } j \\ n_{i,t} & \text{, chosen times for arm } i \end{cases}$$

What if $r_t \in [a, b]$? ($(a, b) \in \mathbb{R}^2$)

- **Multiplicative-MAB**: scaled reward $\hat{q}_{i,t} = \mathcal{S} * \hat{r}_{j,t}$
- **Affine-MAB**: scaled reward $\hat{q}_{i,t}$,

$$\hat{q}_{i,t} = a * \hat{r}_{i,t} + b, \begin{cases} \sum_{i=1}^K \hat{q}_{i,t} = 1 \\ \max_i \hat{q}_{i,t} = \mathcal{S} \end{cases}$$

¹**Finite-time analysis of the multiarmed bandit problem.** P. Auer *et al*
Machine Learning, 47(2/3):235–256, 2002.

D-MABs: Dynamic Multi-Armed Bandits

Environment is dynamic \Leftrightarrow **distributions of rewards change**

How to detect a change in a distribution?

Observations: $\{r_1, r_2, \dots, r_\ell\}$

Q: did the distribution change?

Test: "Page-Hinckley statistics" (with parameters δ and λ)

- 1 $\bar{r}_\ell = \frac{1}{\ell} \sum_{i=1}^{\ell} r_i, m_t = \sum_{\ell=1}^t (r_\ell - \bar{r}_\ell + \delta),$
- 2 $M_t = \max\{m_\ell, \ell = 1 \dots t\}$
- 3 Return ($M_t - m_t > \lambda$)

Dynamic Multi-Armed Bandits: D-MAB = UCB1 + PH



Credit Definition: Scenarios

Uniform Scenario

- $\mathcal{R}_1 = [4..6], \mathcal{R}_2 = [3..5], \mathcal{R}_3 = [2..4], \mathcal{R}_4 = [1..3], \mathcal{R}_5 = [0..2]$
- $\text{credit}(i^{\text{th}}\mathbf{best-op}) \in \mathcal{U}(\mathcal{R}_i)$

from D. Thierens "Adaptive Strategies for Operator Allocation" in *Parameter Setting in Evolutionary Algorithms*, Springer Verlag, 2007.

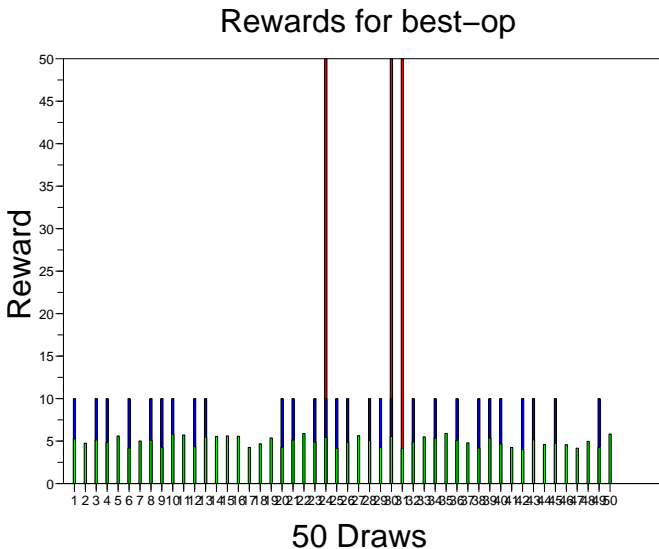
Boolean Scenario

- Probabilities = $\{p_1 = .5, p_2 = .4, p_3 = .3, p_4 = .2, p_5 = .1\}$
- $\text{credit}(i^{\text{th}}\mathbf{best-op}) = 10$ with $p = p_i$, 0 with $p = (1 - p_i)$

Outlier Scenario

- Rewards = $\{r_1 = 50, r_2 = 40, r_3 = 30, r_4 = 20, r_5 = 10\}$
- $\text{credit}(i^{\text{th}}\mathbf{best-op}) = r_i$ with $p = 0.1$, 0 with $p = 0.9$

Credit Definition: Scenarios (II)



Dynamic Rewards Definition

Changes

	best-op	2 nd	3 rd	4 th	worst-op
0	op ₀	op ₁	op ₂	op ₃	op ₄
ΔT	op ₄	op ₁	op ₂	op ₀	op ₃
2 * ΔT	op ₂	op ₄	op ₃	op ₀	op ₁
3 * ΔT	op ₁	op ₂	op ₀	op ₄	op ₃
4 * ΔT	op ₄	op ₁	op ₂	op ₃	op ₀
5 * ΔT	op ₃	op ₁	op ₄	op ₂	op ₀
6 * ΔT	op ₀	op ₄	op ₂	op ₁	op ₃
7 * ΔT	op ₂	op ₃	op ₁	op ₀	op ₄
8 * ΔT	op ₁	op ₄	op ₃	op ₀	op ₂
9 * ΔT	op ₄	op ₀	op ₂	op ₃	op ₁

$$\Delta T \in \{50, 200\}$$

Credit Assignment Rules (ours)

Recall: MAB with UCB

UCB1: at t , choose arm j maximizing $\hat{r}_{j,t} + \sqrt{\frac{2 \log \sum_k n_{k,t}}{n_{j,t}}}$,
 $\hat{r}_{j,t} \in [0, 1]$; for $\hat{r}_{j,t} \in [a, b]$, scaling is needed

MAB with UCB (parameter: \mathcal{S} , for scaling)

- **Multiplicative-MAB:** scaled reward $\hat{q}_{i,t} = \mathcal{S} * \hat{r}_{j,t}$
- **Affine-MAB:** scaled reward $\hat{q}_{i,t}$,

$$\hat{q}_{i,t} = a * \hat{r}_{i,t} + b, \begin{cases} \sum_{i=1}^K \hat{q}_{i,t} = 1 \\ \max_i \hat{q}_{i,t} = \mathcal{S} \end{cases}$$

D-MAB (parameters: \mathcal{S} , for scaling, $\lambda, \delta = 0.15$, for PH test)

- **D-MAB-Multiplicative**
- **D-MAB-Affine**



Probability Matching and Adaptive Pursuit

- $\hat{Q}_{i,t}$ = estimate of reward of arm i , for time t
- At time t , arm i is selected with prob= $s_{i,t}$; gets reward r_t ;

$$\hat{Q}_{i,t+1} = (1 - \alpha)\hat{Q}_{i,t} + \alpha r_t$$

Probability Matching (PM)

$$s_{i,t+1} = p_{min} + (1 - K * p_{min}) \frac{\hat{Q}_{i,t+1}}{\sum_{j=1}^K \hat{Q}_{j,t+1}}$$

Adaptive Pursuit (AP)

$$\begin{aligned} i^* &= \operatorname{argmax}\{\hat{Q}_{i,t}, i = 1 \dots K\} \\ s_{i^*,t+1} &= s_{i^*,t} + \beta(p_{max} - s_{i^*,t}), \\ s_{i,t+1} &= s_{i,t} + \beta(p_{min} - s_{i,t}), \text{ for } i \neq i^* \end{aligned}$$

Parametrization of Methods

Factorial design, 100 runs by parameter setting

- **PM** and **AP**: $p_{min} \in [0, .1]$ (by .01). $\alpha \in [.1, 1]$ (by .1).
- **AP**: $\beta \in [.1, 1]$ (by .1)
- **Mult-MAB**: $\mathcal{S} \in [2, 10]$ (by .1).
- **Affine-MAB**: $\mathcal{S} \in [.1, 2]$ (by .1).
- **D-MAB**: $\delta = .15$, and:
 - Uniform scenario: $\lambda \in [0.5, 10]$ (by 0.5)
 - Other scenarios: $\lambda \in [10, 20]$ (by 2)

Statistical analysis: sequential 1-way ANOVA with confidence 95%, followed by pair-wise comparison (if needed).

Results $\Delta T = 50$ (max. TCR= $2.50 * 10^3$)

TCR: Total Cumulated Rewards; \hat{p}_{best} = prob. choosing best

	AP $P_{\min} / \alpha / \beta$	MAB-M S	D-MAB-M S / λ
Scenario 1: Uniform rewards			
Param.	.03 / .9 / .9	.4	.7/4
TCR	2.12 \pm 0.12	2.25 \pm 0.033	2.24 \pm 0.031
\hat{p}_{best}	0.58 \pm 0.09	0.73 \pm 0.04	0.68 \pm 0.038
Scenario 2: Boolean rewards			
Param.	.03 / .9 / .9	.2	.7 / 21
TCR	1.84 \pm 0.29	1.89 \pm 0.135	1.85 \pm 0.147
\hat{p}_{best}	0.33 \pm 0.13	0.36 \pm 0.05	0.35 \pm 0.07
Scenario 3: Outlier rewards			
Param.	.09 / .1 / \emptyset	.1	.2 / \emptyset
TCR	1.57 \pm 0.49	1.63 \pm 0.24	1.58 \pm 0.234
\hat{p}_{best}	0.22 \pm 0.09	0.24 \pm 0.06	0.22 \pm 0.02

Results $\Delta T = 200$ (max. TCR= $10 * 10^3$)

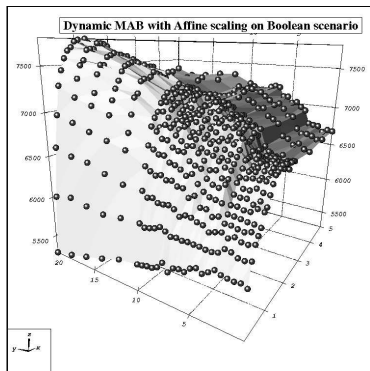
TCR: Total Cumulated Rewards; \hat{p}_{best} = prob. choosing best

	AP $P_{min} / \alpha / \beta$	MAB-M S	D-MAB-M S / λ	D-MAB-A S / λ
Scenario 1: Uniform rewards				
Param.	.02 / .8 / .7	.4	1.2 / 4.5	5 / 8
TCR	9.19 ± 0.25	9.46 ± 0.111	9.75 ± 0.063	9.79 ± 0.059
\hat{p}_{best}	0.75 ± 0.06	0.84 ± 0.009	0.92 ± 0.014	0.94 ± 0.010
Scenario 2: Boolean rewards				
Param.	.02 / .9 / .9	.2	1.5 / 12	1.5 / 16
TCR	8.13 ± 0.38	8.19 ± 0.341	7.57 ± 0.421	7.71 ± 0.347
\hat{p}_{best}	0.48 ± 0.08	0.52 ± 0.05	0.38 ± 0.05	0.41 ± 0.06
Scenario 3: Outlier rewards				
Param.	.07 / .1 / \emptyset	.2	.2 / \emptyset	.3 / \emptyset
TCR	6.66 ± 1.24	7.12 ± 0.585	6.23 ± 0.515	6.17 ± 0.435
\hat{p}_{best}	0.27 ± 0.04	0.29 ± 0.07	0.22 ± 0.01	0.21 ± 0.006

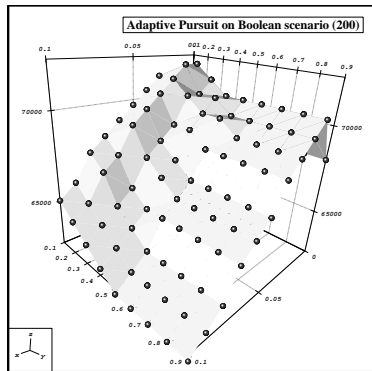
Conclusion - Overall

- **Outlier scenario** poorly handled by all methods
- **MAB**: (surprisingly) performs well in Uniform and Boolean scenarios
- **PM and AP** were competitive... is greediness a good option?
- **D-MAB**: robust with respect to its parameters

D-MAB: robust with respect to its parameters



(a) D-MAB-A



(b) AP

Response surface , *Boolean* scenario ($\Delta T = 200$)

RDV PPSN 2008:

A. Fialho, L. Da Costa, M. Schoenauer and M. Sebag 2008.
Extreme Value Based Adaptive Operator Selection.

Adaptive Operator Selection with Dynamic Multi-Armed Bandits

L. Da Costa¹, A. Fialho², M. Schoenauer^{1,2}, M. Sebag^{1,2}

¹Team TAO, LRI – tao.lri.fr
INRIA Saclay
FRANCE

²Microsoft Research-INRIA Joint Centre
FRANCE

GECCO 2008 – July 16th.